Knowledge Graphs for Enhancing Transparency in Health Data Ecosystems

Maria-Esther Vidal [a,*], Ahmad Sakor [a], Samaneh Jozashoori [a], Emetis Niazmand [a], Disha Purohit [a], Enrique Iglesias [a], Fotis Aisopos [b], Dimitrios Vogiatzis [b], Ernestina Menasalvas [c], Alejandro Rodriguez Gonzalez [c], Guillermo Vigueras [c], Daniel Gomez-Bravo [c], Maria Torrente [d], Roberto Hernández López [d], Mariano Provencio Pulla [d], Athanasios Dalianis [e], Ana Triantafillou [e] and Georgios Paliouras [b]

[a] *Leibniz University of Hannover and L3S Research Center and TIB Leibniz Information Centre for Science and Technology, Germany*
E-mails: maria.vidal@tib.eu, ahmad.sakor@tib.eu, samaneh.jozashoori@tib.eu, disha.purohit@tib.eu, emetis.niazmand@tib.eu, iglesias@l3s.de
[b] *Institute of Informatics & Telecommunications, National Centre for Scientific Research "Demokritos", Greece*
E-mails: fotis.aisopos@iit.demokritos.gr, dimitrv@iit.demokritos.gr, paliourg@iit.demokritos.gr
[c] *Universidad Politécnica de Madrid, Spain*
E-mails: ernestina.menasalvas@upm.es, alejandro.rg@upm.es, guillermo.vigueras@upm.es, daniel.gomez-bravo@upm.es
[d] *Medical Oncology Department, Puerta de Hierro University Hospital, Servicio Madrileño de Salud, Spain*
E-mails: mtorrente80@gmail.com, robertohlopez7@gmail.com, mprovenciop@gmail.com
[e] *Innovation Lab, Athens Technology Center, Greece*
E-mails: T.Dalianis@atc.gr, a.triantafillou@atc.gr

Abstract. Tailoring personalized treatments demands the analysis of a patient's characteristics, which may be scattered over a wide variety of sources. These features include family history, life habits, comorbidities, and potential treatment side effects. Moreover, the analysis of the services visited the most by a patient before a new diagnosis and the type of requested tests, may uncover patterns that contribute to earlier disease detection and treatment effectiveness. Built on the concept of knowledge-driven ecosystems, we devise DE4LungCancer, a data ecosystem of health data sources for lung cancer. Knowledge extracted from heterogeneous sources, e.g., clinical records, scientific publications, and pharmacologic data, is integrated into knowledge graphs. Ontologies describe the meaning of the combined data, and mapping rules enable the declarative definition of the transformation and integration processes. Moreover, DE4LungCancer is assessed in terms of the methods followed for data quality assessment and curation. Lastly, the role of controlled vocabularies and ontologies in health data management is discussed and their impact on transparent knowledge extraction and analytics. This paper presents the lesson learned in the DE4LungCancer development and demonstrates the transparency level supported by the proposed knowledge-driven ecosystem in the context of the lung cancer pilots in the EU H2020 funded project BigMedilytic, the ERA PerMed funded project P4-LUCAT, and the EU H2020 CLARIFY.

Keywords: Healthcare Systems, Data Ecosystems, Knowledge Graphs

*Corresponding author. E-mail: maria.vidal@tib.eu.

1. Introduction

Lung cancer (LC) is the most common cause of cancer death in Europe, with an estimated 353,000 deaths each year. LC has the highest economic cost in Europe, with direct costs of caring for patients with the disease amounting to more than €3 billion per year [51]. Albeit costly, lung cancer therapies can be more effective, and the chances to respond are positively higher when diagnosed in the early stages [16].

Biomedical data have experienced an exponential grown in the last decade; they encode valuable knowledge which can be exploited for accurate disease diagnostics and personalized treatments [46, 48]. Nevertheless, lung cancer is a heterogeneous disease whose precise diagnosis requires a holistic analysis of various variables, usually collected from data sources represented in myriad formats. For example, electronic health records (EHRs) comprise unstructured clinical notes expressed in Spanish, and scientific publications are also unstructured but in English. On the other hand, scientific databases, albeit structured, allow for downloading their collections in semi-structured formats. Natural Language Processing (NLP) techniques are required to recognize biomedical entities and link them to biomedical controlled vocabularies or ontologies in all these data sources. Also, data sharing and processing need to respect data privacy and access regulations imposed by the data providers and ethical and legal committees. Lastly, the decisions made during data processing need to be interpretable and verifiable. These data complexities impose requirements that need to be solved toward a meaningful analysis of the knowledge encoded in these data sources.

Research Goal: The main objective is two-fold: first, we aim to overcome interoperability and data quality issues present in lung cancer data and provide a knowledge-driven framework where analytical methods provide the basis for answering clinical research questions. Second, tasks and decisions implemented in the knowledge-driven framework should be traceable to enhance the framework's transparency and trustability of the analytical results.

Proposed Solution: Built on recent results from the literature [20], we devise a knowledge-driven data ecosystem (DE), named DE4LungCancer, and provide a computational framework to exchange and integrate data while preserving personal data privacy, data security, and ethical and legal regulations. DE4LungCancer is a nested framework that incorporates three DEs: **(i)** Clinical DE: receives unstructured EHRs in Spanish and transforms them into structured databases in tabular (i.e., relational database) and semi-structured (i.e., JSON) formats; **(ii)** Scholarly DE: processes scientific publications related to lung cancer and provides a fine-grained representation of the topics and relations mentioned in a scientific article. **(iii)** Scientific DE: extracts from scientific databases main properties of biomedical entities (e.g., drugs, enzymes, disorders) and their relations or interactions among them (e.g., drug-drug, drug-target, drug-side effect interactions). DE4LungCancer integrates the data processed by each of these DEs and creates a knowledge graph (KG) where data and their meaning coexist. Moreover, the KG comprises entities (modeled as nodes) and their properties and relationships (modeled as edges). Biomedical ontologies and controlled vocabularies are also part of the KG and are utilized to annotate the entities in the KG. These annotations result from the various NLP methods implemented at each basic DE or at the DE4LungCancer DE level. They provide the basis for aligning equivalent entities in the KG. The World Wide Web Consortium (W3C) standard Resource Description Framework (RDF) is used to represent the KG, while the Shapes Constraint Language (SHACL) expresses the integrity constraints over the KG. The KG relies on a unified schema to provide an integrated view of the concepts and properties merged in the KG. The process of data integration is also defined using declarative languages R2RML (a W3C standard), RML (the RDF Mapping Language), and FnO (the Function Ontology). The process of data integration is declaratively defined as mapping rules in RML+FnO; they express correspondences between data sources, and classes and properties from the unified schema. Transformation functions are expressed in FnO and included as part of the mapping rules. This integrated view of data pre-processing and integration results in a modular and reusable specification of the KG creation process, which can be easily verifiable and traceable.

Web APIs have been implemented on top of the KG and over the data processed by each basic DE. The goal is to uncover patterns in the hospital services visited by the lung cancer patients that provide insights about the conditions of these patients before the lung cancer diagnosis. The results of these analyses have driven the design of five clinical interventions to identify which of the hospital services visited by lung cancer patients, have more potential of diagnosis, and may contribute to earlier detection. The reported results uncover patterns in the visited services that provide insights about the potential clinical conditions of patients diagnosed with lung cancer. Although further analyses are required, these patterns can support early diagnosis and prognosis. More importantly, if validated, they

will allow clinicians to detect the disease in the asymptomatic phase, reducing complications, which usually increase the complexity of these patients and their outcome.

DE4LungCancer has been applied in the context BigMedilytics[1], P4-LUCAT[2], and EU H2020 CLARIFY [3]. BigMedilytics is a European Union Horizon 2020 funded project that aims at developing innovative data-driven solutions to improve the healthcare system in Europe. BigMedilytics covers a wide range of chronic diseases and frequent cancers (e.g., prostate, lung, and breast). Specifically, in the lung cancer pilot, the goal is to process biomedical data sources and uncover patterns that enhance the understanding of the risk of suffering lung cancer or the effectiveness of treatment. Data sources are in different formats. P4-LUCAT is an ERA-NET project in the area of Personalized Medicine to support oncologists in the prescription of lung cancer treatments. CLARIFY is a European Union Horizon 2020 research and innovation project, funded to exploit biomedical data and Artificial Intelligence techniques to identify risk factors that may deteriorate a patient's condition after oncological treatment. CLARIFY covers lymphoma, and lung and breast cancer. DE4LungCancer is integrated into the whole data ecosystem of the CLARIFY framework to enable the management of lung cancer clinical records from the Puerta del Hierro University Hospital in Madrid. In these projects, DE4LungCancer enables the integration of biomedical data sources and provides a knowledge graph from where analytical methods are performed. As a proof of concept, this paper presents some of these methods, and reports on the outcomes that have motivated the execution of clinical interventions to enhance treatment effectiveness and lung cancer patients' quality of life.

This article comprises six additional sections. Section 2 presents requirements to be satisfied at data management, clinical, and ethical and legal level in the context of lung cancer. DE4LungCancer is defined in section 3, and section 4 describes the data quality issues assessed in the data ecosystems that compose DE4LungCancer. Section 5 presents the evaluation of DE4LungCancer, and section 6 summarizes the state of the art. Lastly, section 7 wraps up and outlines future work.

2. Challenges in Health Data Ecosystems

European Health Data Ecosystems target at strengthening the sustainability of health systems across Europe by reducing costs while improving quality and access to care [4]. They aim at broadening the repertoire of computational methods that bolster conscientious diagnosis and treatments. Moreover, they require the definition of best practices for data sharing and integration, as well as for preserving privacy and ethical regulations.

2.1. Requirements in Health Data Ecosystems

Requirements can be classified in three categories: **Data Management:** include all the needs to be satisfied during sharing, curation, management, processing, and analysis of both data and metadata. **Clinical:** the requirements in this category correspond to requests stated by the oncologists to support the design of clinical interventions. **Ethical & Legal:** this category comprises the requirements for preserving personal data privacy and security and ensuring that ethical and legal regulations are fulfilled.

Data Management Requirements (DRs). The requirements in this group are aligned with the needs for data management in data ecosystems proposed by Geisler and Vidal et al. [20]. **DR1** Management of data in various formats, e.g., unstructured clinical records and scientific publications, semi-structured data in scientific databases like DrugBank. **DR2** Data and metadata must satisfy the integrity constraints defined by clinicians. Moreover, all the decisions made for data quality assessment and data curation need to be interpretable and verifiable. **DR3** The processes of data management need to be transparent. In addition, stakeholders should be able to trace the steps implemented

[1] https://www.bigmedilytics.eu/
[2] https://p4-lucat.eu/
[3] https://www.clarify2020.eu/
[4] https://digital-strategy.ec.europa.eu/en/policies/strategy-data#:~:text\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{C\global\mathchardef\accent@spacefactor\spacefactor}\let\begingroup\def{}\endgroup\relax\let\ignorespaces\relax\accent9C\egroup\spacefactor\accent@spacefactorommon%20European%20data%20spaces%20will,improve%20health%20care

to transform data from different formats and integrate them into a unified knowledge base (a.k.a. knowledge graph). Specifically, to satisfy **DR3**, we consider **DR4** and state that the data sources should be defined in terms of a unified schema that describe all the properties of the entities collected in the data sources. The correspondences or mappings among data sources and the unified schema should be declarative and available to be checked and verified.

Clinical Requirements. These requirements are specified in terms of key performance indicators (KPIs). As a proof of concept, we present five of the KPIs that have guided the development of DE4LungCancer; however, the techniques presented in this paper are generic and can be applied to satisfy other KPIs. The aim of these five KPIs is to discover the factors that impact the patients' quality of life and the usage of healthcare services. The data ecosystem should offer services on top of the integrated data to check these KPIs. The validation of these KPIs is through medical interventions with the goal of optimizing them. **KPI1:** Duration in days of the hospital stays of the lung cancer patients. **KPI2:** Identification of patients at risk of developing lung-cancer. **KPI3:** Number of admissions to the emergency room in a given time period. **KPI4:** Toxicities observed in lung cancer patients who suffer from comorbidities and receive oncological and non-oncological drugs. **KPI5:** Degree of satisfaction of the lung cancer patients treated by oncologists supported by the DE4LungCancer services.

Ethical Requirements (ERs). The requests in this category are also aligned with the Ethical and Legal requirements proposed by Geisler and Vidal et al. [20], the European Union guidelines for Trustworthy AI [17], and the regulations of the Spanish Law of Personal Data Access [5] (Leyes 15/1999 and 41/2002). **ER1** Follow a legal framework where patient privacy is respected, and clinical records are utilized as indicated in the consent granted by the lung cancer patients. **ER2** Accounting bias and fairness to guarantee that none of the recommendations given by data ecosystem analytical tools is affected by sensitive attributes (e.g., age or ethnicity). **ER3** Traceability of the satisfaction of data privacy regulations during the whole process of data ingestion, processing, integration, and analysis. **ER4** Documenting and explaining quality issues.

2.2. A Lung Cancer Data Ecosystem

The main goal of a lung cancer data ecosystem is to develop analytical tools that give the oncologists insights to improve the management of patients with lung cancer during their treatment, follow-up, and last period of life through data-driven techniques. Additionally, they aim to improve patients' experience, satisfaction, and primary outcomes and save substantial health costs. Moreover, admissions and readmissions due to toxicities and comorbidities present in lung cancer patients need to be traced to reduce visits to emergency care and hospitalizations. A lung cancer data ecosystem should provide the basis to identify the potential side effects of a lung cancer treatment and the adverse events generated by the interactions among the treatment drugs.

There are four different categories of stakeholders in a Lung Cancer data ecosystem; data are exchanged across these stakeholders, preserving data access and privacy regulations. **Oncologists:** clinical partners who are responsible for treating the lung cancer patients, collecting the clinical data, defining the clinical goals, and designing and running the clinical interventions. **Computer Scientistics:** technical partners who develop all the techniques to ingest, process, integrate, and analyze the pilot data. They are also in charge of devising all the methods to preserve data privacy and respect what is stated in the patients' consents. **Ethical & Legal Boards:** experts in ethical and legal regulations to preserve data privacy. **Software Developers:** technical partners who develop the computational framework and implement the data ecosystem.

3. DE4LungCancer

The DE4LungCancer framework is devised as a network of data ecosystems (DEs)[20]; it aligns data and metadata to describe the network and its components. Heterogeneity issues across the different datasets are overcome by various methods of data curation and integration. Each DE comprises datasets, and programs for accessing, managing, and analyzing their data. Interoperability issues across the datasets of the DEs are solved in a unified schema. Mappings between the datasets and the unified schema describe the meaning of the datasets. Figure 1 illustrates the

[5] https://www.boe.es/buscar/act.php?id=BOE-A-2002-22188

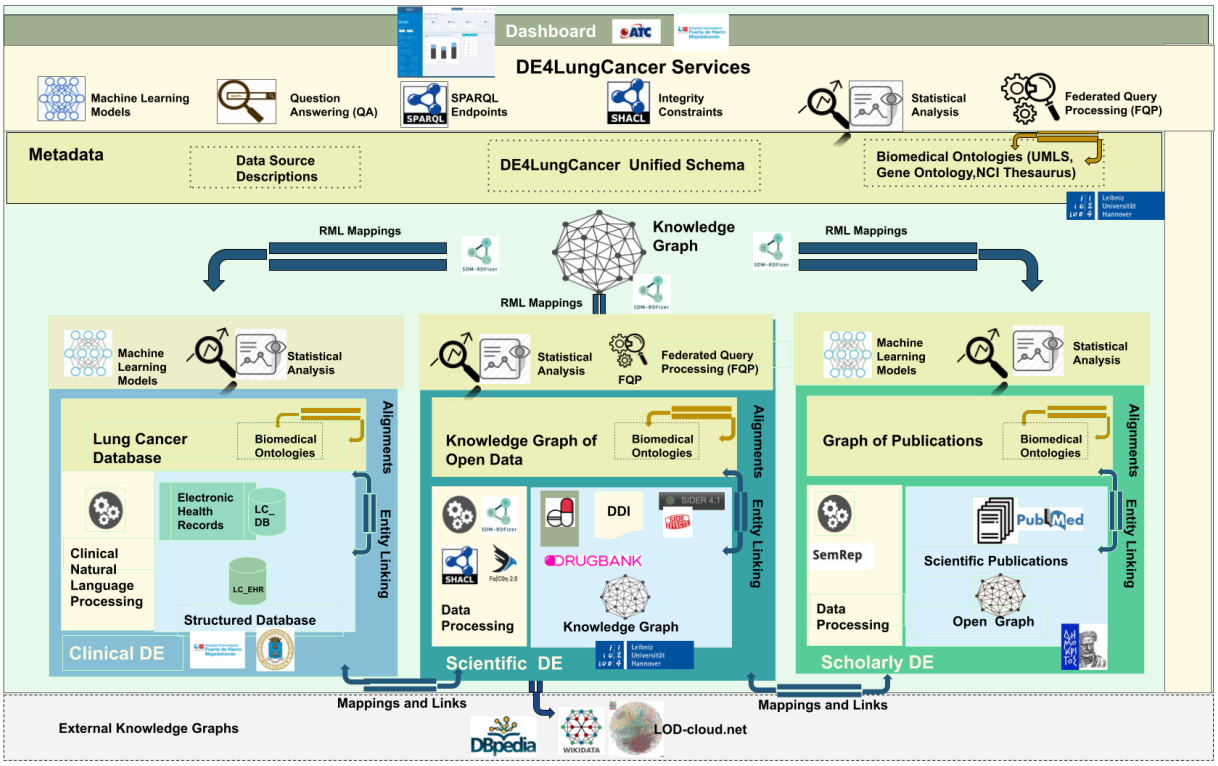


Fig. 1. The DE4LungCancer Data Ecosystem

components of the DE4LungCancer Data Ecosystem. The metadata layer specifies biomedical vocabularies (e.g., Unified Medical Language System-UMLS[6] or Human Phenotype Ontology-HPO [7]). The DE4LungCancer DE is a nested framework which is also composed of three basic DEs: Clinical, Scholarly, and Scientific Open. Each basic DE can also comprise datasets, metadata, and programs.

3.1. Basic Data Ecosystems

3.1.1. Clinical Data Ecosystem

Clinical data is collected from electronic health records from more than 1,300 lung cancer patients registered in the Electronic Health Record (EHR) system at the Puerta del Hierro University Hospital in Madrid from 2008 till January 2020. The data is extracted from 315,891 notes and 16,550 reports and represents clinical variables of lung cancer patients and the services consulted by those patients before and after diagnosis. The data is anonymized. The statistical analysis performed on EHR follows a stage of Natural Language Processing of raw data to extract patient characteristics, and visited medical services at the hospital. The (statistical) analysis performed on EHR concerned KPI-1: Length of hospital stay; and KPI-2: Identification of people at risk of developing lung-cancer. Raw data: 988 EHR of patients from 2008-2020, 416 patients were hospitalized 942 times. From these 416 patients, 166 had one hospitalization in the first three months after diagnosis. The remaining 250 did not have a hospitalization in the first three months after diagnosis, but they had at least one hospitalization up to six months after the diagnosis. NLP processing on EHR, technical details: Natural Language Processing (NLP) techniques are applied to EHR to extract relevant entities from unstructured fields, i.e., clinical notes or lab test results. The NLP techniques rely on medical vocabularies, e.g., UMLS or HPO, and on an NLP library, to perform lemmatization, Named Entity Recognition

[6] https://www.nlm.nih.gov/research/umls/index.html
[7] https://hpo.jax.org/app/

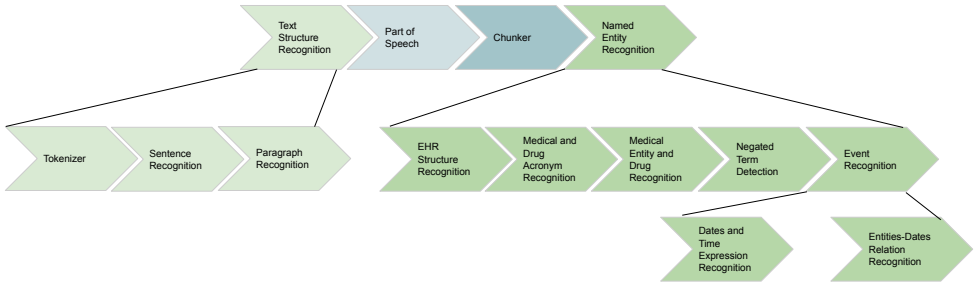


Fig. 2. NLP Pipeline for Knowledge Extraction in the Clinical Data Ecosystem

(NER). The final result is an annotation of the extracted concepts (i.e., Named Entities) with terms from medical vocabularies. Figure 2 depicts the NLP pipeline that transforms unstructured EHR into a structured database.

3.1.2. Scholarly Data Ecosystem

Scholarly data are obtained by harvesting scientific publications from PubMed (i.e., article abstracts) and PubMed Central (i.e., article full-texts), along with publication scholarly metadata such as the author list, journal, and publication year. In order to retrieve publications only related to lung cancer, the Entrez Programming Utilities API available by PubMed[8] is used with the MeSH topic "Lung Neoplasms", and also collect the rest of the MeSH topics related to each article. Moreover, except from the scholarly metadata available in PubMed, we also retrieve other type of information such as the citations of each publication, by querying the Scopus Citations Count API [9], as well as the Hirsch index (h-index) and SCImago Journal Rank (SJR) indicator of each journal, available from the SCImago Lab[10] in the context of the Scimago Journal & Country Rank project[11]. Natural Language Processing is also applied to the article abstracts, as well as on the whole text of the article if it is freely available. Out of the aforementioned resources, it produces triplets that are made of two entities connected by a relation (e.g., Hemofiltration-TREATS-Patients). The process is performed with industry standard software. Metamap [12] performs named entity recognition, returning the named entities and a confidence factor. SemRep[13] performs relation extraction while relying on MetaMap. The result of the NLP is the open graph. The NLP pipeline is depicted in Figure 3. In addition, the original articles are annotated with MeSH terms, and UMLS annotates the resulting triplets. This data mining task has many

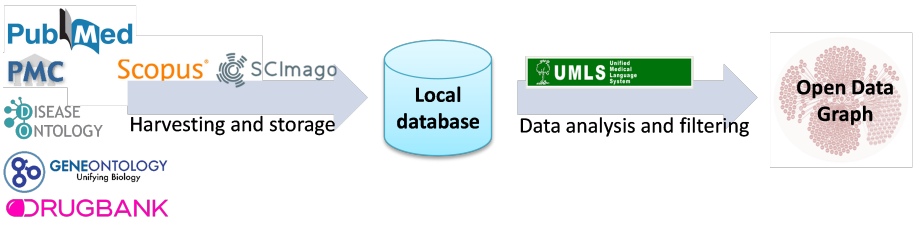


Fig. 3. NLP Pipeline for Knowledge Extraction Implemented in the Scholarly Data Ecosystem

potential applications to the current pilot, including but not limited to drug-drug interaction discovery. In specific,

[8] https://www.ncbi.nlm.nih.gov/home/develop/api/
[9] https://dev.elsevier.com/documentation/AbstractCitationCountAPI.wadl
[10] http://www.scimagolab.com/
[11] https://www.scimagojr.com/
[12] https://metamap.nlm.nih.gov/
[13] https://semrep.nlm.nih.gov/

we utilize the open graph to address the problem of predicting new drug-drug interactions as a binary classification problem for interacting/non-interacting drug pairs. To this end, we employ a machine learning technique analyzing the undirected semantic paths connecting different pairs of drugs in the open graph [7]. The sequence of relations in semantic paths are used to create a set of features for a big number of drug pairs related to lung cancer. Those features are then used to train a Random Forest classifier that is able to effectively discriminate between two classes: interacting and non-interacting pairs, using Drugbank as a gold standard. The classifier was able to produce 22,346 drug-drug interactions predictions with a certain confidence score, based on the drug pairs' semantic paths. We also conduct a comparative study of existing node embedding methods for biological network mining tasks. Apart from tensor factorization methods, we also utilize deep learning techniques. Moreover, we apply these approaches to ten different real-world biological datasets and evaluate their performance on three downstream mining tasks: i) link prediction, ii) node classification, and iii) clustering. We extend the analysis by i) including state-of-the-art deep learning methods, ii) adding more real-word datasets, and iii) performing more mining tasks. Results show that node embedding methods can be effectively used in all three tasks. In addition, comparison between state-of-the-art approaches helps choose the appropriate method with respect to the task. Matrix and tensor factorization approaches perform better in link prediction tasks as well as in multilayered networks. On the other hand, deep learning methods achieve better results on the node classification and clustering tasks. Our evaluation is performed on ten public real-world datasets. We used seven multi-layered protein-protein interaction networks (h1v, plasmodium, xenopus, saccpomb, rattus, bos, celegans), a drug-drug interaction network (ddi), a single protein-protein interaction network (ppi with labels), and a medical term-term relation network (cooc with labels).

3.1.3. Scientific Open Data Ecosystem

Scientific databases (e.g., DrugBank and SIDER) and encyclopedic knowledge bases (e.g., Wikidata and DBpedia) are the main sources of open data. These sources encode knowledge about drugs, the approved indications of these drugs, the side effects, and the drug-drug interactions and their effects. All these features are present as short textual descriptions. In the pilot, we have collected the following data from DrugBank (version 5.1.8 in XML) and SIDER (2018 tabular). We have downloaded 1,550,586 drug-drug interactions, 60,177 side effects of drugs, 2,333 drug indications, and 10,150 drug-target and 4,523 drug-enzyme interactions. These collections of data enable the understanding of the impact on toxicities and drug effectiveness of the drugs of an oncological treatment. The techniques of named entity recognition (NER) and named entity linking (NEL) enable the identification of biomedical entities from textual attributes. The rule-based entity linking engine, FALCON [44], is used to perform NER and NEL on this data ecosystem. FALCON is configurable for linking entities to diverse controlled vocabularies or knowledge graphs (KG), e.g., UMLS, DBpedia, or Bio2RDF. FALCON recognizes entities by mapping instances of a word within a short text, i.e., a surface form, into the textual representation of entities in the controlled vocabulary or KG. FALCON resorts to a knowledge base and a catalog of rules for recognizing and linking entities. The knowledge base integrates various sources, e.g., DBpedia, Wikidata, Oxford Dictionary, and Wordnet. Additionally, it comprises alignments between nouns and entities in these sources. Alignments are stored in a text search engine, e.g., ElasticSearch, while the knowledge sources are maintained in an RDF triple store accessible via SPARQL endpoints. Moreover, the catalog of rules encodes the English morphology; they are represented as conjunctive rules and provide a forward chaining inference process for entity recognition in English short texts. The main feature of FALCON is the ability of splitting an input short text into the minimal number of entities that more precisely represent the words in the text. Thus, FALCON is devised to solve the optimization problem of maximizing the number of words linked to an entity/relation while minimizing the number of recognized entities/relations. This feature is extremely relevant for the scientific open data ecosystem entity, e.g., a drug or a disease, can be expressed with several words, e.g., mepregenol diacetate or thoracic aortic aneurysms.

3.2. The Nested DE4LungCancer Data Ecosystem

As proposed by Geisler and Vidal et al. [20], the nested DE4LungCancer Data Ecosystem is defined as a 6-tuple *DE*=⟨*Datasets, Data Operators, MetaData, Mappings, Integrity Constraints, Services*⟩.

3.2.1. Datasets

The DE4LungCancer Data Ecosystem integrates three categories of data sources collected from the basic data ecosystems: ***Processed Clinical Data*** Database produced by the Clinical Data DE as the result of the EHR NLP; 1,042 EHRs are described in terms of 320 attributes. The data is structured and presented as a nested structure in JSON. Additionally, the information about the hospital services visited by the patients is shared in a relational database. The values of the attributes are in English and Spanish, and attributes like treatments are diagnostics are annotated with terms from UMLS. ***Scholarly Data*** A data graph– in Neo4J [14]– representing 162,394 scientific publications in a graph with 402,020 nodes and 12,256,983 edges. Each publication is described with a PubMed identifier, title, year, journal, authors, SCImago Journal rank indicator (sjr), Hindex, number of citations, and the link to SCOPUS with all the information of the article. Moreover, publications are annotated with 4,821,501 associations describing the relationship *has topic*, 7,368,157 associations for the relationship *mention in*, and 166,219 associations between UMLS terms. ***Scientific Open Data*** 11,292 drugs described in terms of the conditions for which the drug can be prescribed, and its interactions with targets and enzymes. There are also 60,177 relations between drugs and side effects, 1,550,586 drug-drug interactions extracted from the Literature and DrugBank, and 502,839 predicted drug-drug interactions discovered by various predictive methods.

3.2.2. Metadata

Biomedical ontologies and controlled vocabularies are utilized to describe the data and provide a unified description and annotation. These annotations represent the basis of the data integration methods followed to merge the data into a KG. The values in the datasets are annotated with terms from the Unified Medical Language System (UMLS) [15]. These annotations enable entity alignment and provide the basis for the integration of the datasets into the KG. The KG includes 3,862,429 terms from the semantic groups "Anatomy", "Disorders", "Physiology", "Procedures", "Concepts & Ideas", "Chemicals & Drugs", "Living Beings", "Activities & Behaviors", "Objects", "Devices", "Phenomena", "Occupations", "Organizations", "Geographic Areas", and "Genes & Molecular Sequences". A unified schema provides an integrated view of the data sources. The DE4LungCancer unified schema is expressed in the W3C standard data model RDF. This increases interoperability and facilitates reusability of existing vocabularies and ontologies, e.g., the RDF Schema[16] (RDFS), the Web Ontology Language[17] (OWL), and PROV-O[18] (Provenance Ontology). The unified schema composes 80 classes, 64 object properties, and 110 datatypes. To ensure findability and availability, the unified schema is published [19] in Vocol [23] at the TIB-Leibniz Information Centre for Science and Technology. VoCol is a collaborative platform for ontology development that enables the development of vocabularies using Version Control Systems. VoCol brings the following advantages: **Collaborative Support:** Several users can work simultaneously in the development of the ontology, and changes are synchronized automatically. **Quality Assurance:** Syntactic validation of the unified schema to be compliant with RDF, RDFS, and OWL, and semantic validation for consistency checking. **Analysis:** VoCol provides ontology management features that enable the visualization and exploration of the ontology. VoCol also provides an interface for specifying queries against the unified schema, and its classes and properties. Finally, the documentation describing the metadata of each class and property can be consulted, as well as basis analysis summarizing the number of classes and properties that compose the unified schema. Thus, FAIR principles (Findability, Availability, Interoperability, and Reusability) [21] are respected; they represent the basis of a transparent plan for data management in DE4LungCancer.

3.2.3. Mappings

The correspondences between the data sources and the unified schema are defined using the W3C standards RDF Mapping Language (RML) [12] and R2RML. R2RML and RML mapping rules can comprise transformation functions expressed in existing ontologies (e.g., the Function Ontology-FnO). These mappings are expressed in RDF and can be stored into a triplestore (e.g., Virtuoso or GraphDB). Listing 1 presents a SPARQL query that collects

[14] https://neo4j.com/
[15] https://www.nlm.nih.gov/research/umls/index.html
[16] https://www.w3.org/TR/rdf-schema/
[17] https://www.w3.org/TR/owl-features/
[18] https://www.w3.org/TR/prov-o/
[19] http://ontology.tib.eu/bigmedilytics/

```
PREFIX rr:  <http://www.w3.org/ns/r2rml#>
PREFIX rml: <http://semweb.mmlab.be/ns/rml#>
PREFIX bm:  <http://bigmedilytics.eu/vocab/>

SELECT DISTINCT ?mappingRule ?logicalSource ?predicate ?sourceAttribute
WHERE {
?mappingRule rml:logicalSource ?ls.
?ls          rml:source        ?logicalSource.
?mappingRule rr:subjectMap     ?subject.
?subject     rr:class          bm:LCPatient.
OPTIONAL {   ?mappingRule rr:predicateObjectMap ?pObjectMap .
             ?pObjectMap  rr:predicate          ?predicate .
             ?pObjectMap  rr:objectMap          ?objectMap .
             ?objectMap   ?mode                 ?sourceAttribute}}
```

Listing 1: SPARQL Query to Retrieve RML Mapping Rules defining the class LCPatient

```
PREFIX rr:   <http://www.w3.org/ns/r2rml#>
PREFIX rml:  <http://semweb.mmlab.be/ns/rml#>
PREFIX fnml: <http://semweb.mmlab.be/ns/fnml#>
SELECT DISTINCT ?functionCall ?argument ?action  ?argumentValue
WHERE {
?functionCall fnml:functionValue    ?fv.
?fv           rr:predicateObjectMap ?pom .
?pom          rr:predicate          ?argument .
?pom          rr:objectMap          ?om .
?om           ?action               ?argumentValue
FILTER (?action in (rml:reference, rr:constant))}
```

Listing 2: SPARQL Query to Retrieve FnO Functions Called in Mapping Rules

the information about the mapping rules that define the class `bm:LCPatient`. The projected attributes include the data source from where the data is collected, and per predicate of the class, the attribute(s) of the corresponding data source used to populate the predicate. Moreover, Listing 2 depicts a SPARQL query that retrieves the functions called from the mapping rules. The projected attributes include a function call, the function arguments, the action performed over the arguments, and the value of the argument. These functions are expressed in FnO and are part of the toolbox *EABlock*[20] [27, 28]. This toolbox includes functions that solve entity alignment over biomedical textual attributes. They are built on top of FALCON [44] for solving the tasks of Named Entity Recognition (NER) and Entity Linking (EL). Specifically, three functions are used; they enhance data quality by aligning the recognized biomedical entities to terms in UMLS, Wikipedia [49], and DBpedia [4]. More importantly, the specification in RDF and the semantic description using FnO provides a standard documentation of entity alignment and establishes the basis for tracking down the process of the data integration. In the DE4LungCancer DE, the combination of R2RML, RML, and FnO represent a powerful formalism to specify the pipeline for integrating data into the KG declaratively. Moreover, as observed in Listings 1 and 2, this specification enhances transparency and facilitates the traceability of the decisions taken during KG creation.

The DE4LungCancer KG is defined in terms of 524 RML mappings that include 20 calls to five of *EABlock* func-

[20] https://zenodo.org/record/5779773#.Ym1FC_MzZTY

tions. A SPARQL endpoint with the unified schema and the triples maps is publicly available [21]; the execution of the SPARQL queries in Listings 1 and 2 provides a view of the metadata that describes the management processes implemented on top of the datasets.

The mapping rules have been defined by two knowledge engineers and reviewed by another two knowledge engineers, clinicians, and technical partners. These rules have been devised taking into account the concepts in the DE4LungCancer unified schema, the metadata describing the concepts in the DE4LungCancer data sources, and communications with the domain experts. Figure 4 reports on the number of RML mapping rules per class and their properties in the unified schema. In average, a class is defined by 9,4 mapping rules (standard deviation 16,4). In particular, `bmLC:Annotation` and `bmLC:LCPatient` are defined using 116 and 40 mappings, respectively.

SDM-RDFizer [25], an in-house RML compliant engine, is utilized to integrate data from the data sources into the KG following the mapping rules. As a result, a KG of 19,602,972 biomedical entities described in terms of 110,788,660 RDF triples is created. Moreover, 3,900,764 links to DBpedia, Wikidata, and UMLS are part of the KG; they are discovered by the tasks of NER and NEL executed by the FnO function included in the mapping rules and by the NLP processes implemented in each DE. Figure 4 depicts the number of entities of the classes in the KG. The classes `bmLC:MENTION_IN`, and `bmLC:HAS_TOPIC` are populated with entities extracted from scientific publications, while `bmLC:Annotation` comprises the UMLS terms that annotate the entities recognized by the NER implemented on top of the DE4LungCancer datasets. Figure 6 depicts a portion of the DE4LungCancer KG for an entity representing an anonymous lung cancer patient (a.k.a. `bmLC:LCPatient`). As shown, an `bmLC:LCPatient` entity is directly associated with properties that include lung cancer stage, performance status, oncological and non-oncological treatments, visited hospital services, observed toxicities, and surgeries. Moreover, through the UMLS annotations, an `bmLC:LCPatient` entity is connected to **(i)** Publications whose topics are the values of the entity properties; **(ii)** Drug-drug interactions–reported or predicted– of the patient's prescribed treatments; **(iii)** Effects of the patient's prescribed treatment reported in the publications or scientific databases (e.g., DrugBank). The amount of data and knowledge associated with each of an `bmLC:LCPatient` entity provides a

[21] https://labs.tib.eu/sdm/bm-mappings/sparql

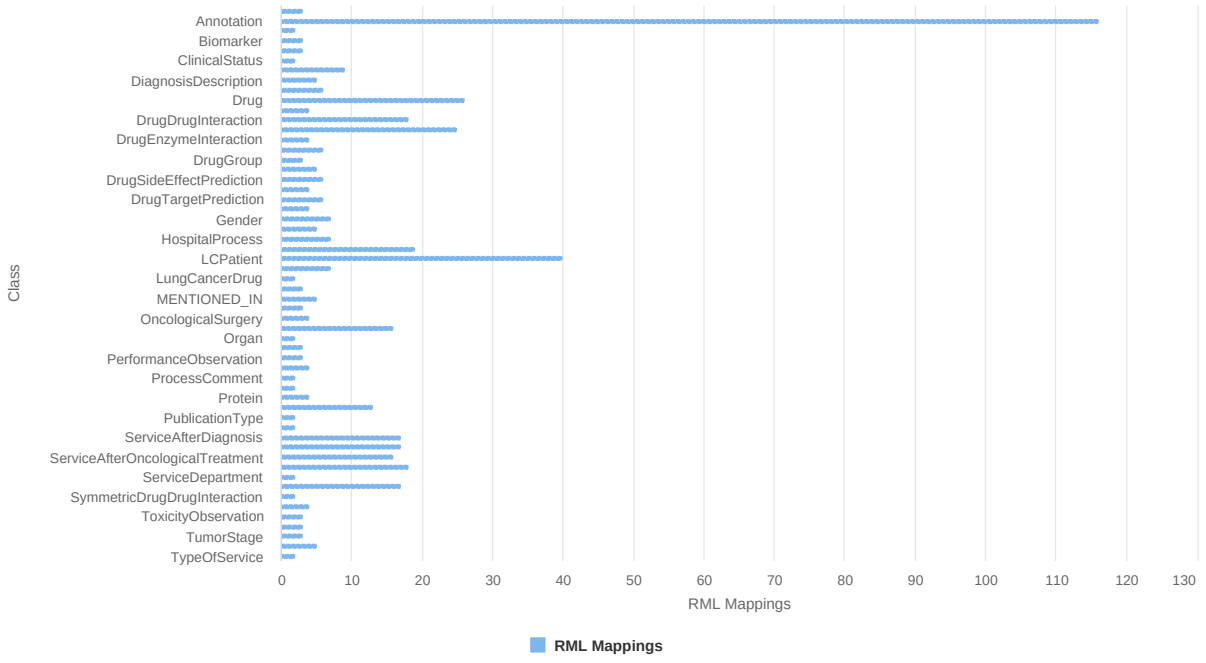
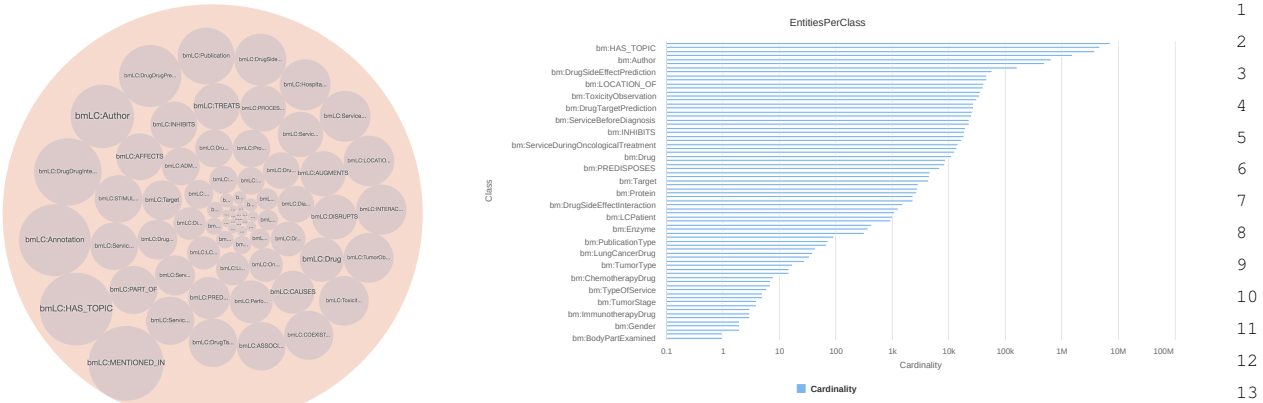


Fig. 4. RML Mapping Rules Per RDF Class in the DE4LungCancer unified schema



(a) KG Classes

(b) Entities per Class

Fig. 5. RDF Classes in the KG. Annotations of entities extracted from Publications (`bmLC:MENTION_IN` and `bmLC:HAS_TOPIC`), Clinical Records and Scientific Open Data (`bmLC:Annotation`) are the most populated. `bmLC:MENTION_IN`, `bmLC:HAS_TOPIC`, and `bmLC:Annotation` have 7,368,157; 4,821,501; and 3,862,414 entities.

holistic profile of `bmLC:LCPatient` entities. The knowledge encoded in these profiles is extracted from scientific publications and databases, as well as modelled in the unified schema and annotations from controlled vocabularies (e.g., UMLS). As a result, these profiles enhance interpretability of the clinical records and provide the foundations for explainable predictions and discoveries. More importantly, these profiles make easy the validation of the KPIs in an integrated fashion.

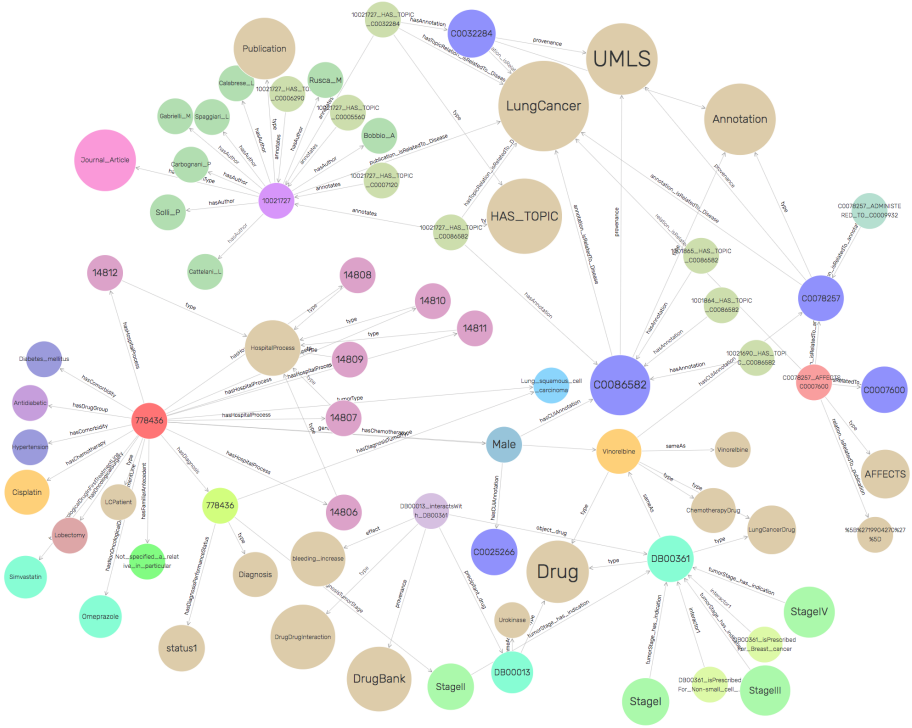


Fig. 6. A Portion of the DE4LungCancer KG

3.2.4. Integrity Constraints

The integrity constraints to be satisfied by the data sources are expressed in terms of rules which are validated with the clinical partners to ensure completeness and soundness. To enhance traceability, the integrity constraints are expressed declaratively using the Shapes Constraint Language (SHACL) [30] and SPARQL [39]. SHACL rules are defined over a class's attributes (i.e., owl:DatatypeProperties), and constraints on incoming/outgoing arcs, cardinalities, RDF syntax and extension mechanisms. Inter-class constraints induce a shape network used to validate the integrity and data quality properties of the KG. SHACL relies on validation errors and the validation reports using a controlled vocabulary in RDF. A validation report includes explanations about the violations, the severity of the violation, and a message describing the violation. In total, there have been defined 67 rules to validate the integrity of the DE4LungCancer KG.

3.2.5. Services

The mapping rules are also utilized to validate the correctness of the process of KG creation. In the DE4LungCancer DE, a quality assessment process guided by the mapping rules is executed on top of the KG. It ensures that each class and predicate in the KG has the same number of instances as the data sources from where the data to populate these classes and properties are extracted. Moreover, different services have been implemented on top of the KG. Several statistical analyses derived from various parameters (e.g., hospitalization, emergency room visits, toxicities, medical tests performed, and oncological treatment types) are integrated as services of the DE4LungCancer dashboard. Additionally, services to traverse the scientific publications associated with a cohort of patients or the drug-drug interactions and side effects of these patients' treatments can be explored. Thus, the KG acts as a knowledge repository of the DE4LungCancer DE which empowers interpretability of the conditions and treatments of the selected cohort. Lastly, DE4LungCancer resorts to a federated query engine to interoperate across the DE4LungCancer KG, DBpedia, and Wikidata; the query processing methods proposed by Endris et al. [13] are implemented to ensure that data privacy regulations are respected during the execution of queries against the DE4LungCancer KG.

Table 1
Number of Constraints and Ambiguities in the Class bmLC:LCPatient

Class	# Constraints	# Ambiguities Detected
Biomarker	8	134
Smoking Habit	4	47
Vital Status	3	0
Familial Antecedents	2	0
Oncological Surgery	5	303
Biopsy	3	52
Performance Status	4	75
Tumor Stage	8	1,486
Oncological Drug	6	177
Oncological Treatment Line	24	846
Total	67	3,120

4. Data Quality (DQ) and Ethics in the DE4LungCancer Data Ecosystem

4.1. DQ in Clinical Data Ecosystem

The data quality methodology is composed of four steps: (a) Definition of the constraints; (b) Validation of the constraints; (c) Human validation by the domain experts; and (d) Resolution of the ambiguities. First, the metadata describing the DE4LungCancer data sources as well as the description of the universe of discourse represented in these data sources are analyzed with the aim of identifying integrity constraints. Clinical partners and technical partners were consulted in order to collect the main constraints to be satisfied. Moreover, the concepts and relations existing in the unified schema were utilized to guide the definition of the constraints. First, constraints describing

the properties of the attributes of a class in the DE4LungCancer unified schema were identified, i.e., *intra-concept* constraints, and next, constraints regarding the relationships existing between these concepts or *inter-concept constraints* are determined. *Intra-concept constraints* include (a) data types of the attributes, (b) attribute dependencies, (c) cardinalities, and functional dependencies. Additionally, inter-concept constraints encompass referential integrity, cardinality and connectivity, and mandatory and optional relationships among the concepts in the unified schema. Once the constraints are recognized, they are formally specified as expressions of SQL, SHACL, and SPARQL, and evaluated both over the corresponding raw data and the data integrated in the KG; the SHACL validation engine Trav-SHACL [18] was used to validate the SHACL constraints against the KG. Moreover, inconsistencies between the results obtained after the evaluation of the constraints over raw data and the KG, reveal errors in the process of integration in the KG. On the other hand, equal numbers of ambiguities in the raw data and the KG evidence a data quality issue in the original dataset or in the extraction process. Finally, when all the issues have been detected and classified, the clinical and technical partners were consulted to find the most suitable way to curate either the raw data or the KG. This methodology implements techniques reported by Acosta et al. [2], Ruckhaus et al. [43], and Mihaila et al. [35]. For the class LCPatient, all the attributes were analyzed, as well as the concepts to which this concept is connected. Table 1 summarizes *intra-* and *inter-concept constraints* in the business domain of lung cancer. These constraints have been validated by four knowledge engineers, two experts in the NLP extraction process, and two experts in lung cancer; all these evaluators are partners of the consortium. As a result, 67 constraints are defined and a total of 3.120 ambiguities are detected in the NLP processed clinical datasets and in their corresponding instances in the KG. Table 1 reports on distribution of the constraints attributes and concepts. As observed, the majority of the ambiguities are detected in the tumor stages, line of treatments, oncological surgeries, and biomarkers. All these ambiguities were discussed with the clinical partners and curated following their recommendations and directions. The integrity constraints are part of the metadata of the DE4LungCancer DE; they document the quality assessment and curation tasks, and trace the changes made during data curation.

4.2. DQ in the Scholarly Data Ecosystem

In contrast to clinical data that have been manually filled by experts, the knowledge published in scholarly sources may usually be less reliable. Although being reviewed by the field experts, published literature can still report preliminary results, observations and unverified hypotheses. Moreover, given that any NLP software used to automatically extract knowledge from text is far from perfect, we expect a significant amount of inherent noise and unreliable information in the open graph (i.e., the one resulting from processing of the scientific publications). As mentioned, we employ two mainstream tools in the field of biomedical knowledge extraction, in order to perform entity recognition and relation extraction on literature text. MetaMap [3] and SemRep [41] tools have been evaluated on benchmark datasets achieving high precision (>76%) and moderate recall (36% - 70%), on various datasets [9, 11, 29]. The quality of data in the open graph produced by those tools is addressed in two ways. First, each triplet is associate with a quality score, that is related to the confidence scores provided by MetaMap, representing the quality of each concept identification. In specific, a triple-extraction quality score ranging from 0 to 1 (i.e., the higher, the better) has been added, by averaging the concept identification score of the subject and object entities of each triple. These concept identification scores provide the average of the scores for all found instances of the entities in the specific relation, in order to take into consideration the frequency of the concepts found in the scientific publications.

Second, in order to assess the quality of the open graph as a whole, we have developed an error detection methodology [8] that is based on graph topology and theoretic measures to assess the quality of all edges in this graph. This method, called Path Ranking Guided Embeddings (PRGE), combines an extension of the Path Ranking Algorithm [31] (PaTyBRED [34]) with translational graph embeddings (TransE [6]). The aim is to generate confidence-guided graph embeddings that will identify erroneous triples, by providing global-confidence scores for all automatically generated relations. The PRGE has been evaluated using two benchmarks and one generated dataset, with its AUC score ranging from 0.56 to 0.97, based on the quality of dataset used, and the noise imputation approach followed each time, improving in most cases simple PRA and embedding methods.

Apart from the errors imputed by automatic NLP systems, the quality of information provided by publications can also be dubious. From the end user perspective, when exploring publications included in the Scholarly DE, it would be appropriate to be able to filter out unreliable publications or focus only on trustworthy institutes and journals. To

Table 2
Number of UMLS Annotations

DE4LungCancer Class	UMLS
bmLC:CAUSES	8,699
bmLC:PREDISPOSES	7,017
bmLC:ADMINISTERED_TO	2,910
bmLC:ASSOCIATED_WITH	27,948
bmLC:DISRUPTS	12,828
bmLC:TREATS	18,820
bmLC:INTERACTS_WITH	40,331
bmLC:MANIFESTATION_OF	319
bmLC:LOCATION_OF	45,523
bmLC:PROCESS_OF	17,005
bmLC:AUGMENTS	13,873
bmLC:HAS_TOPIC	4,821,501
bmLC:COEXISTS_WITH	31,549
bmLC:STIMULATES	23,317
bmLC:INHIBITS	19,260
bmLC:AFFECTS	19,849
bmLC:MENTIONED_IN	7,368,157
bmLC:PART_OF	25,486
Total Distinct Links	12,485,564

this end, DE4LungCancer provides with the ability to explore scientific literature, using various factors as filters for the information retrieved:

- Journal: Different journals have different standards in the review process and the completeness of the published work. As a measure of quality of each journal, we provide the journal h-index, as well as the SCImago Journal Rank (SJR) indicator.
- Authors: An expert can be interested in publications by universities or specific authors that can be known for their overall contribution in the field. Thus, filtering can be applied by author name or affiliation.
- Publication type: Different types of articles are defined according to the different levels of evidence (e.g., scientific review or clinical trial) based on which the represented knowledge is derived. Accordingly, the type of publication is also provided to allow for relevant filtering.
- Publication year: The age of a publication allow an expert to decide if the results depicted in the publication are up-to-date. Therefore, the publication year, as another useful filter for the end-users, is also provided.
- Cited By Count: The number of citations for a specific publication can provide a good indication of its quality and trustworthiness.

Table 2 reports the number of annotations from UMLS extracted by the Natural Language Processing techniques implemented in the Scholarly Data Ecosystem. These DE4LungCancer classes correspond to relations in the UMLS Semantic Network [22]. The entities corresponding to scientific publications have been annotated with 12,485,564 terms from UMLS. Together with the ones extracted by the Scientific Open DE (Section 4.3), these annotations establish the entity alignments required for the data integration process in the DE4LungCancer KG. Section 4.4 presents the effects of including the links in all the biomedical entities that populate the DE4LungCancer KG.

4.3. DQ in Scientific Open Data Ecosystem

Out of 1,550,586 drug-drug interactions (DDI) collected from DrugBank, 320 patterns were recognized in order to evaluate the performance of FALCON in this use case, 1,198 DDI descriptions were manually annotated by twelve annotators; annotations correspond to CUIs from UMLS and constitute the gold standard of the evaluation. For example, for the DDI description: "The serum concentration of Lepirudin can be decreased when it is combined with Tipranavir"; Lepirudin and Tipranavir correspond to the extracted entities from the above record, while decrease and

[22] https://www.ncbi.nlm.nih.gov/books/NBK9679/

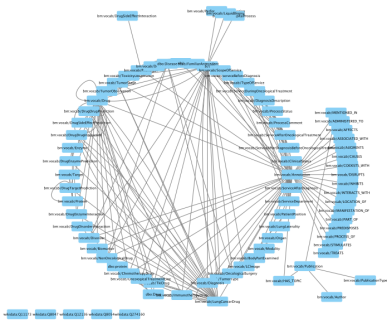
serum concentration represent, respectively, the effect and impact of the interaction of Tipranavir with Lepirudin. A 2-fold cross-validation was followed while building the gold standard, and disagreement was solved by a majority voting. The evaluation indicates a precision of 98%. The 2% where FALCON failed to extract and link the terms correctly, are interactions which contain more than one interaction in the same sentence, where FALCON was only considering one interaction. Additionally, the *EABlock* toolbox has been assessed in *Baseline* and *EABlock* pipelines. Three sets of RML mapping rules were evaluated on two datasets of biomedical concepts composed of 10K and 20K entities, respectively. Both pipelines generate the same KGs. However, in *Baseline*, NER and NEL are performed in a pre-processing stage of KG creation, while *EABlock* functions were executed with the RML rules in the *EABlock* pipeline. Observed execution time suggests that using the *EABlock* functions speeds up the KG creation process by up to 40%. Moreover, we created five gold standard datasets considering textual values with frequent quality issues that frequently exist in textual values datasets, (e.g., character capitalization, elimination, insertion, and replacement). These datasets are built from DBpedia, Wikidata, and UMLS. The errors are introduced with certain percentage of the records (i.e., 50% and 80%). The *EABlock* functions exhibited a F1 score that varied from 0.78 in DBpedia, 0.88 in UMLS, and 0.99 in Wikidata.

Table 3
Number of Links From the DE4LungCancer KG to UMLS, DBpedia, and Wikidata

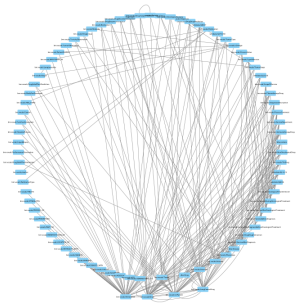
DE4LungCancer Class	UMLS	DBpedia	Wikidata
bmLC:BodyPartExamined	1	0	0
bmLC:LungLaterality	2	0	0
bmLC:ProcessStatus	4	0	0
bmLC:FamilialAntecedent	7	7	0
bmLC:Biomarker	3	0	0
bmLC:Gender	2	2	0
bmLC:Phenotype	77	76	61
bmLC:Enzyme	373	368	353
bmLC:Disorder	435	425	403
bmLC:ClinicalStatus	2	0	0
bmLC:OncologicalSurgery	5	5	0
bmLC:Modality	2	0	0
bmLC:PatientPosition	1	0	0
bmLC:Target	4,364	4,299	4,180
bmLC:ImmunotherapyDrug	3	3	0
bmLC:TypeOfService	3	0	0
bmLC:ServiceDepartment	43	0	0
bmLC:TumorStage	4	0	0
bmLC:ProcessComment	3	0	0
bmLC:NonOncologicalDrug	44	43	0
bmLC:LungCancerDrug	39	38	2
bmLC:Diagnosis	5	0	0
bmLC:TkiDrug	7	7	0
bmLC:Drug	7,323	6,457	3,175
bmLC:DiagnosisDescription	273	0	0
bmLC:TumorType	17	17	0
bmLC:ChemotherapyDrug	8	8	0
Total Distinct Links	12,961	11,679	8,172

Table 3 reports on the number of links recognized by *EABlock* during the execution of the RML+FnO mapping rules that define the DE4LungCancer KG. In total, 12,961, 11,679, and 8,172 distinct DE4LungCancer entities are connected to UMLS, DBpedia, and Wikidata[23]. These links have been manually curated by the DE4LungCancer knowledge engineers. These results indicate that the name entity recognizer and linker used in *EABlock* toolbox (i.e., FALCON) is not able to completely recognize and link medical terms, and exhibit better performance in UMLS and

[23]Note that an entity (e.g., a drug) may belong to various DE4LungCancer classes, this explains why the total numbers of links do not correspond to the sum of the number of links of each individual DE4LungCancer class.



(a) $G_{NoLinks}$ without Entity Alignment



(b) G_{Links} with Entity Alignment

Analysis	G_{Links}	$G_{NoLinks}$
Number of nodes	82	82
Number of edges	216	180
Avg. number of neighbors	4.9	4.0
Clustering coefficient	0.13	0.09
Network density	0.03	0.02
Connected components	1	6

(c) Network Analysis

Fig. 7. Network analysis to assess connectivity of $KG_{NoLinks}$ and KG_{Links}. Aggregated graphs $G_{NoLinks}$ and G_{Links} represent provide a summarized view of the number of connections in $KG_{NoLinks}$ and KG_{Links}.

DBpedia than in Wikidata. They corroborate the outcomes reported by Sakor et al. [45] and indicate that further research is required to enhance the completeness of the named entity linking task over Wikidata.

4.4. DQ in the Nested DE4LungCancer Data Ecosystem

The annotations from UMLS, DBpedia, and Wikidata extracted by the NLP techniques implemented in DE4LungCancer enable to create the entity alignments that define the semantic data integration process executed during the creation of the DE4LungCancer KG. This section presents the impact that these annotations have in semantic data integration. This impact is measured in terms of connectivity or number of alignments that they enable to establish in the DE4LungCancer KG.

Two versions of KGs are created: $KG_{NoLinks}$ and KG_{Links}, the latter includes the links discovered by NER and NEL tasks executed in the RML+FnO mapping rules, while in the former these links have not been generated. The links are removed from the classes `Drug`, `Enzyme`, `Indication`, `Target`, and `Toxicity`. These KGs are aggregated into two directed graphs $G_{NoLinks}$=$(V, E_{NoLinks})$ and G_{Links}=(V, E_{Links}). Vertices in V keeps the classes in $KG_{NoLinks}$ and KG_{Links} with at least one entity. A labelled directed edge e=(q,p,k) belongs to $E_{NoLinks}$ (resp., to E_{Links}) if there are classes Q and K in V, and q and k are instances of Q and K in V, and the RDF triple (q, p, k) belongs to $E_{NoLinks}$ (resp., E_{Links}). Traditional network analysis methods are conducted on top of $G_{NoLinks}$ and G_{Links} to determine connectivity. The metrics are (a) The Average number of neighbors indicates the average connectivity of a vertex or node in a graph. (b) Clustering coefficient measures the tendency of nodes who share the same connections in a graph to become connected. If a neighborhood is fully connected, the clustering coefficient is 1.0 while a value close to 0.0 means that there is no connection in the neighborhood. (c) Network density measures the portion of potential edges in a graph that are actually edges; a value close to 1.0 indicates that the graph is fully connected. (d) The number of connected components indicates the number of subgraphs composed of vertices connected by at least one path. Figure 7 depicts the aggregated graphs $G_{NoLinks}$ and G_{Links}, and Figure 7c reports on the results of the graph metrics. The outcomes indicate that KG_{Links} comprises more connected entities. Albeit low, the clustering coefficient and density values indicate that the UMLS annotations and links to DBpedia and Wikidata included in KG_{Links}, increase the connectivity. As a result, these connections allow for the integration into the DE4LungCancer KG of the biomedical entities annotated individually in each of the data ecosystems that composed the DE4LungCancer framework. Moreover, based on the results reported by Waagmeester et al. [15], which put Wikidata into perspective as a knowledge graph for the life sciences, the recognized links enrich the DE4LungCancer KG with the richness of knowledge collected and maintained by the scientific communities in DBpedia and Wikidata.

5. The DE4LungCancer Assesment

5.1. Data Management Requirements

The data management techniques implemented in DE4LungCancer enable to overcome the data management requirements: DR1-Data variety; DR2-Integrity constraint satisfaction; DR3-Transparent data management; and DR4-Unified definition of heterogeneous data. The definition of the KG creation process using declarative languages (e.g., R2RML, RML, and FnO) imposes a significant workload on the team of knowledge engineers. They should understand the data sources and unified schema, as well as the interoperability issues existing across the data sources. Despite the cost of human workload, the declarative definition of data integration brings significant benefits. They empower the reusability and modularity of the data integration process. More importantly, this process facilitates the traceability of the decisions made to integrate raw data into entities of the KG, and to curate data quality issues, and enhance interpretability of the detected data quality issues.

5.2. Quantitative Analysis of Top of the DE4LungCancer KG

A dashboard makes the DE4LungCancer KG available to the clinical partners in the lung cancer pilots of BigMedilytics and CLARIFY, and in P4-LUCAT. Various services are provided to analyze the processed EHRs and the holistic profiles that integrate EHRs with the fine-grained representation of publications and scientific open data. The dashboards are available to the project oncologists via certificate-based authentication. The outcomes of the analytical tools provided by the dashboard have established the basis for the implementation of clinical interventions for the lung cancer patients treated by the team of oncologists of the Puerta del Hierro University Hospital in Madrid. The improvement of the diagnostic pathway and the reduction in the length of hospital stays and emergency rooms represent key points that have been analyzed. With this aim, the DE4LungCancer KG can be traversed to identify the most visited services by patients with a new diagnosis of lung cancer in the previous 15 months to diagnosis; in this analysis, four months before diagnosis to avoid consultations related to the diagnostic process strictly, such as medical oncology. Moreover, the services explored to identify the prescribed clinical tests.

We retrieved all the properties of 1,051 patients from the DE4LungCancer KG; 859 patients visited at least one first-attention service between the day of the diagnosis and 15 months prior to this date. 459 patients saw first-attention services between four and 15 months previous to diagnosis; 331 were in stage III or IV of lung cancer. During the month in front of the diagnosis of lung cancer, which we used as our baseline, the most visited services were: Thoracic surgery, Pneumology, Medical Oncology, Internal Medicine, and Emergency Room. When analyzing 15 months ahead of lung cancer diagnosis, excluding the four months before diagnosis, the top-5 most visited services are General Emergencies, Primary Care, Cardiology, Pneumology, and General and Digestive Surgery. Additionally, we have observed that patients have an increase in the number of first-attention consultations during the 15 months prior to diagnosis of lung cancer. Moreover, the visited services differ entirely from those seen during the month prior to diagnosis of lung cancer. The number of tests is also increased during this period.

The hospital services, visited for the first time by the lung cancer patients, are grouped into six categories according to the number of months prior to the lung cancer diagnosis. These groups are denoted as `X-Y` indicating that the group includes all the health services visited, the first time, by a lung cancer patient during the months *Y* prior to the lung cancer diagnosis but excluding the period between the day of the diagnosis and the month *X* before the diagnosis; i.e., `0-1` includes all the health services visited by a lung cancer patient during the month previous to the lung cancer diagnosis. Figure 8 shows the evolution of the top-10 most visited services in the periods `0-1`, `0-4`, `4-12`, `4-13`, `4-14`, and `4-15`. General Emergencies and Primary Care are the services more frequently visited in the periods `4-13`, `4-14`, and `4-15`. Spearman's Rho (SR), a non-parametric test used to measure the correlation between two ordered sets, is used to compare the services most visited in different periods prior to diagnosis. Figure 9a and Figure 9b present heatmaps reporting on Spearman's Rho and p-value, respectively; the average value of Spearman's Rho is 0.64 and the average p-value is 0.097. These results suggest that `4-14` and `4-15` are the most stable periods in terms of frequency of patients visiting, as first attention, the hospital services (i.e., the Spearman Rho Index value is 0.87 with a p-value of 0.0012).

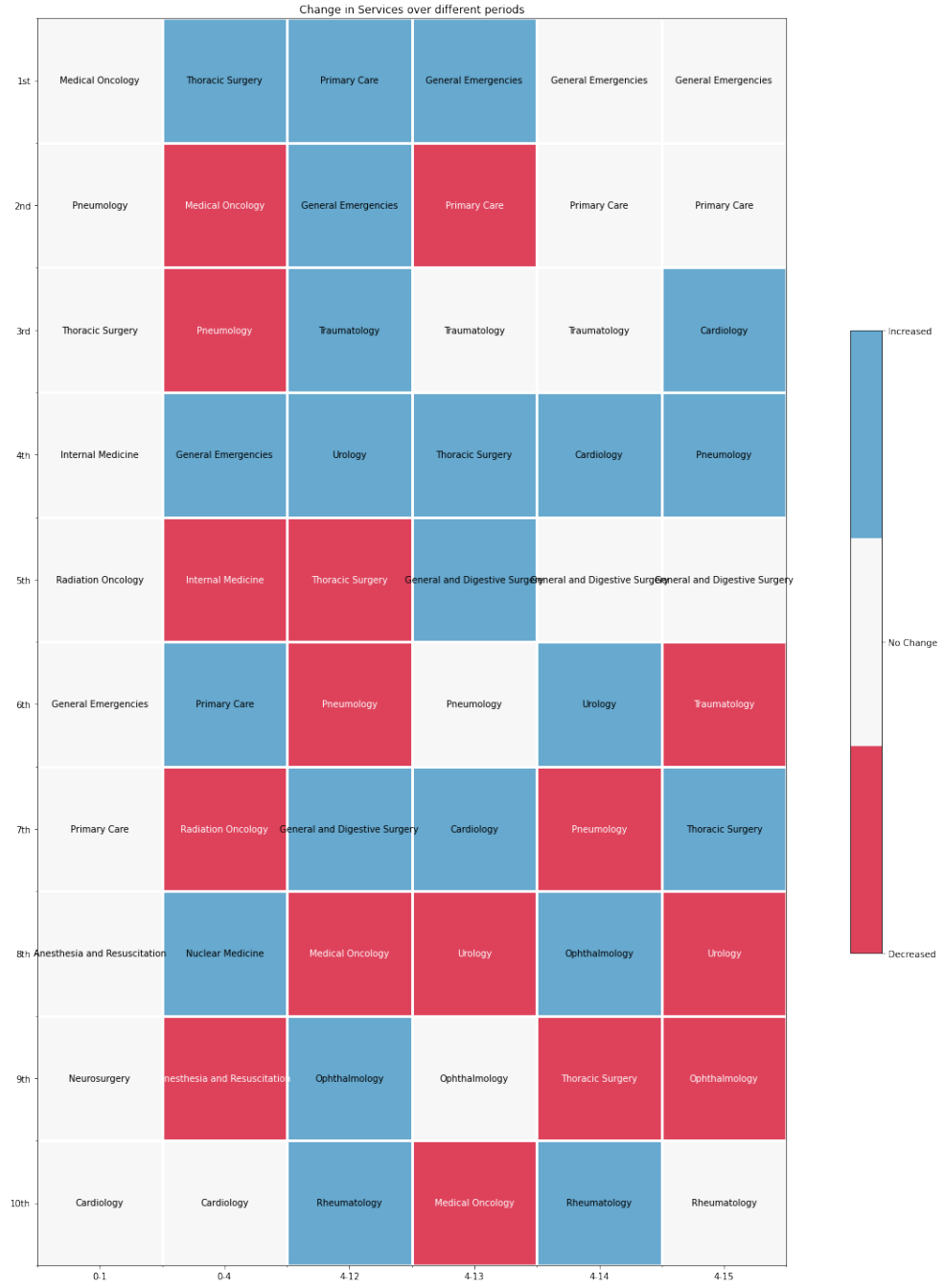
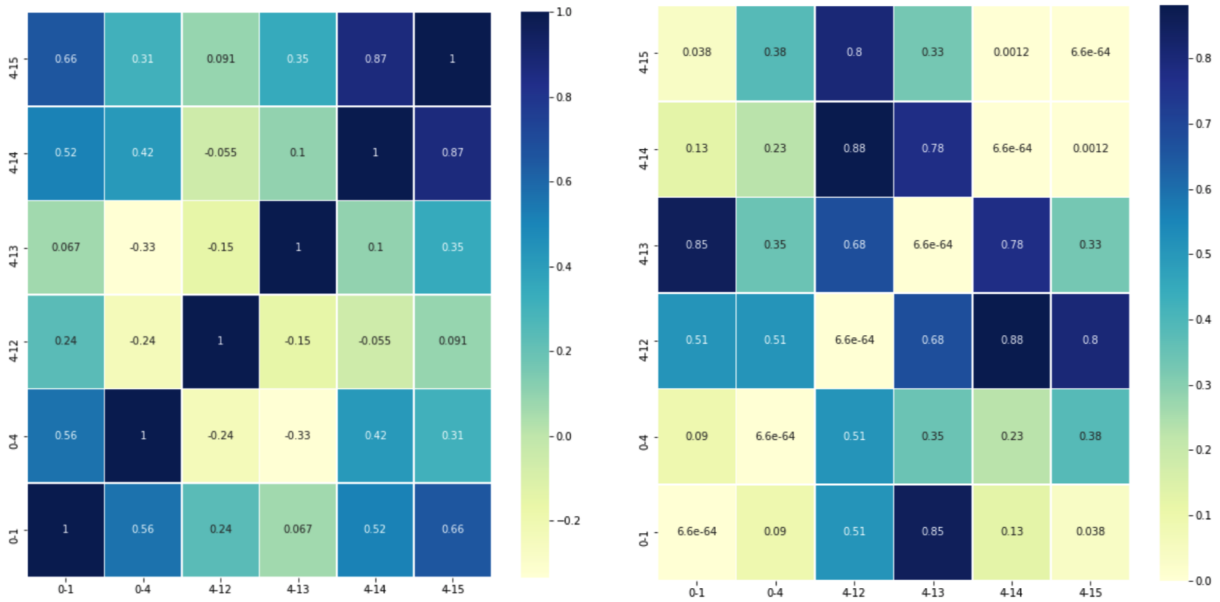


Fig. 8. Evolution of the Top-10 hospital services most visited the first time by the lung cancer patients prior to the lung cancer diagnosis. Blue indicates that the number of visits of the service increases, and it moves up in the list. Red shows that the number of visits of the service decreases, and it moves down in the list. White shows a service position stays the same with the respect to the previous reported period.

Additionally, we compute the Jaccard index to quantify the overlap between sets of services visited in distinct periods; Figure 10 reports these results. The average Jaccard index is 0.62, indicating a relatively high overlap across the studied periods. However, the first attention services visited one month and four months before the diagnosis are



(a) Spearman Rho index

(b) P-values of the Spearman Rho index values

Fig. 9. Comparison of the most visited hospital services in the periods `0-1`, `0-4`, `4-12`, `4-13`, `4-14`, and `4-15`

the same (i.e., Jaccard Index is 1.0). Furthermore, the first attention services visited in the periods `4-14`, and `4-15` are almost the same (i.e., Jaccard Index is 0.82), but differ from the services visited in the other studied periods.

These results uncover patterns in the visited services that provide insights about the potential clinical conditions of the patients diagnosed with lung cancer. Although further analysis is required, these patterns can support early diagnosis and prognosis. More importantly, if validated, they will allow for detecting the disease in the asymptomatic phase, reducing complications, which usually increase the complexity of these patients and their outcome.

5.3. Clinical Interventions

Based on the results of the quantitative analysis conducted on top of the DE4LungCancer, the clinical partners devised five interventions; they assess the satisfaction of the DE4LungCancer clinical requirements KPI1, KPI2, KPI3, KPI4, and KPI5. These interventions aim at studying a lung cancer patient at various stages of the lung cancer pathway, i.e., healthy person, person with persistent medical conditions, sick person, person in treatment, and terminally ill person (Figure 11).

1st intervention- Identification of people at risk of developinng lung cancer: The goal is the identification of people at risk of developing lung cancer and a continuous assessment of a patient's bypass channels. This intervention has been possible, speeding up appointments for diagnostic tests as well as consultation reviews when it comes to a patient with suspected cancer.

2nd intervention- Characterization of the Lung Patients: Based on the analysis results of the services most frequently visited by patients in the 15 months prior to diagnosis, first-attention visits to certain services (e.g., General Emergencies, Primary Care, Cardiology, and Pneumology) are considered relevant patterns. As a result, persons who follow these patterns are selected as patients, who may be in an asymptomatic stage and may have the potential risk of developing cancer.

3rd intervention- Identification of the Most Visited Hospital Services During Lung Cancer Followed-Up: General Emergencies has been identified as the most visited medical service once a patient is under follow-up by the Oncology Department. Pain is one of the most common symptoms, because pain is often changed with disease progression. Despite the importance of pain assessment and management, it is uncovered that pain

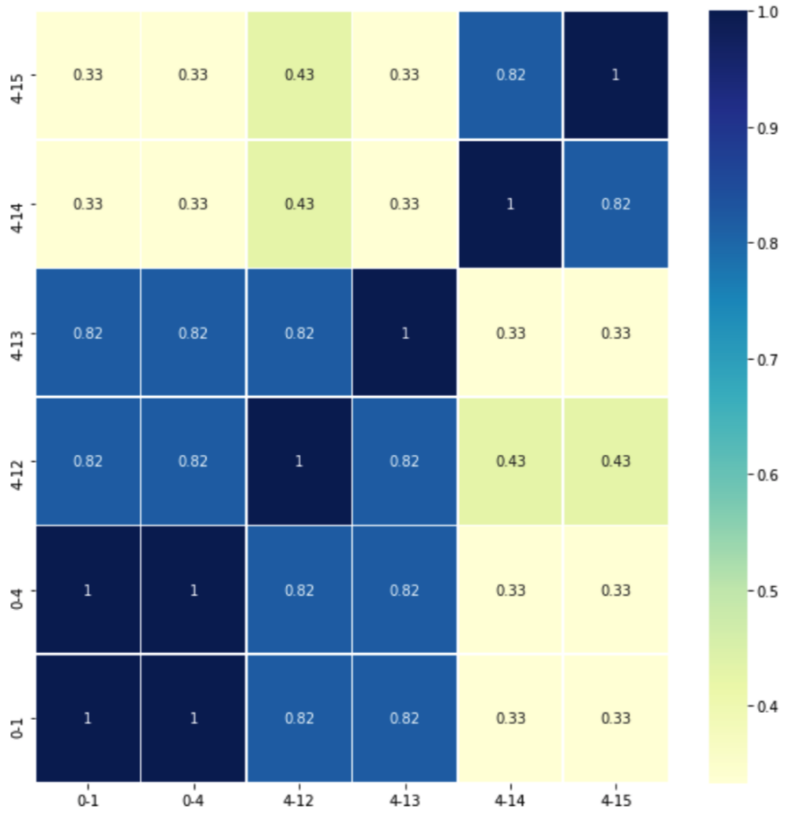


Fig. 10. Jaccard Index Values. Overlap of most visited hospital services in the periods `0-1`, `0-4`, `4-12`, `4-13`, `4-14`, and `4-15`

under treatment is common. Thus, this intervention aims at reinforcing the work of the nurses in assessing pain and favoring early referral to the Pain Unit.

4th intervention- Study of Combination of Comorbidities and Specific Treatments: This intervention is defined based on the analysis of the patients who attended General Emergencies and were readmitted, in a new hospital service, in a period of 28 days. The study aims at uncovering combinations of comorbidities and specific treatments that increase the risk of being readmitted to the emergency room. Based on the uncovered patterns, the Oncology Department processes inter-consultations to the departments of the most visited hospital services to identify potential side effects of the prescribed treatments.

5th intervention- Administration of Palliative Attention and Provide Close Control: The goal of the intervention is to administrate Palliative Care attention and provide close control in consultations before the next treatment cycle date in a treatment line. This study has allowed for measuring readmission and death at 28 days after discharge to determine the need for external early clinical control. Furthermore, the frequency of this event in advance or initial treatment lines has been assessed. This quantitative analysis indicates that 30% of the patients are over 70 years old; they also suffered advanced stages of lung cancer and more than three comorbidities. Additionally, they have received more than three lines of oncological treatments. These results are considered as a pattern to promote lung cancer patients to palliative care.

5.4. Ethical & Legal Requirements

Data sharing, management, and analysis in the DE4LungCancer DE have been conducted following the regulations imposed by Ethical protocol and the Ethical committee of the Puerta del Hierro University Hospital in Madrid. Thus, a legal framework to respect a data privacy has established (Requirement **ER1**). Following these regulations,

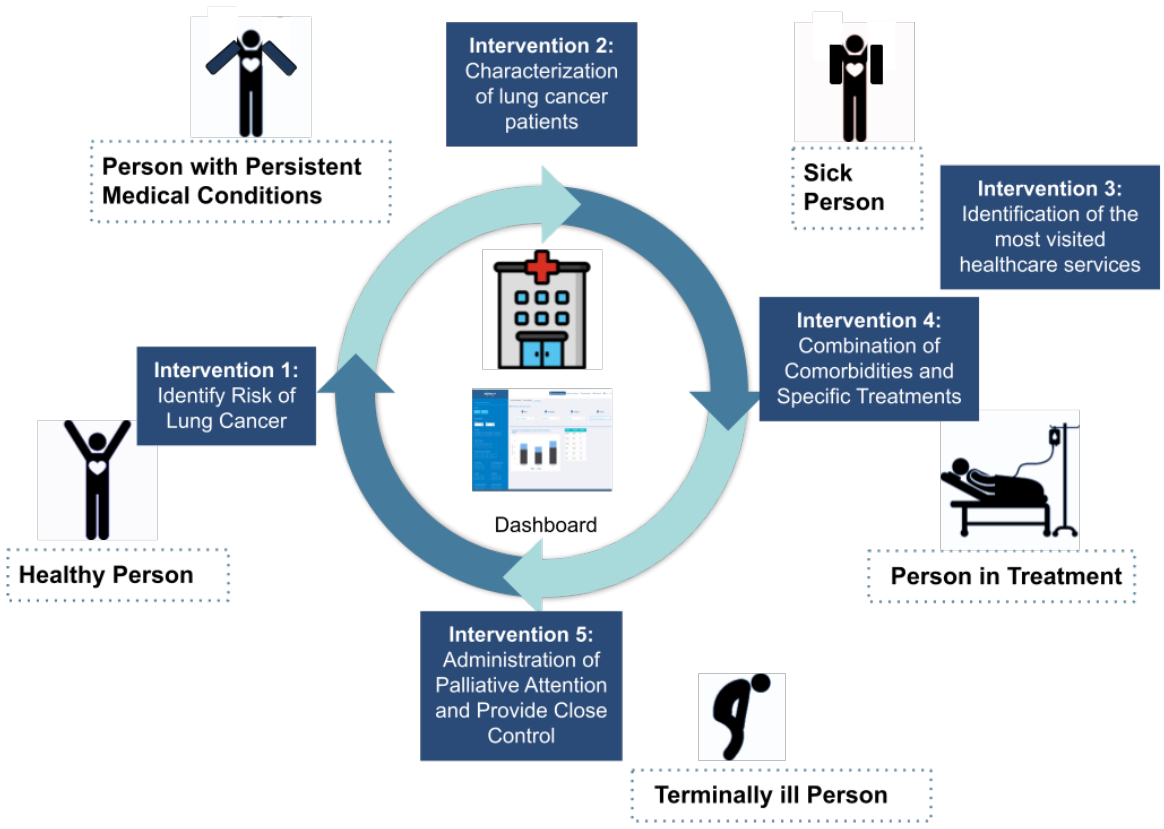


Fig. 11. Lung Cancer Pathway and Medical Interventions

anonymization techniques have been implemented at the hospital and the DEs have been conducted only services according to the patients' consents. The consent is in a writing document signed by each patient. It states the purpose of the research and the funding agency, the reasons to participate in the study, and the possible advantages and disadvantages of being part of the study. Taking into account national laws, the organizations that share and manage clinical data (raw and processed) have established the conditions to comply with the data protection requirements of the project in the areas of: risk assessment, data protection impact assessment, data protection by design, security measures, privacy notice, data sharing and processing agreement, and record of processing activities. Sensitive attributes have been removed from the processed data integrated in the KG (Requirement **ER2**). These tasks have been evaluated and validated by the data protection officer of these institutions (Requirement **ER3**). Lastly, every detected ambiguity in the data that can be considered as a data quality issue has been documented and verified with the clinical partners; all the decisions for data curation are documented (Requirement **ER4**).

6. Related Work

Health Data Management: Medical big data analyses suffers from various technical issues (missing values, dimensionality etc.) and bias control [32]. Various initiatives aim to address the nexus of technical, legal, ethics-related, governance and data protection-related, and cultural challenges arising for health data ecosystems [33]. A first attempt to document and analyze all research, regulatory and ethical requirements stemming from the aggregation and analysis of clinical data from different health organizations was presented in [50], providing a solution ar-

chitecture considering technical and organizational aspects. Other works [10, 47] focus on the ethical and regulatory challenges that surround AI in healthcare, analyzing the data protection and privacy requirements that will ensure fairness and transparency of related approaches. DE4LungCancer resorts to Semantic Web technologies (i.e., RML, FnO, SHACL, and UMLS) to enhance transparency during data integration and knowledge graph creation.

Big Data Ecosystems: A Data Ecosystem can be defined as a complex socio-technical network, enabling collaboration between autonomous actors in order to explore and analyze big data [38]. Efficient, transparent and ethical data management is an ultimate goal in such projects utilizing big data. This becomes more evident when dealing with biomedical or clinical data that are characterized by high sensitivity [1, 36] and noisy entries [53]. DE4LungCancer is built on the knowledge-driven framework devised by Geisler and Vidal et al. [20] and makes available a network of nested data ecosystems able to exchange semantically described metadata. This knowledge enrichment empowers DE4LungCancer with transparency and facilitates traceability of data integration and analytics.

Biomedical Knowledge Graphs: Knowledge graphs have gained attention as expressive data structures that enable the convergence of the data and knowledge using a graph data model [22]. They provide a common understanding of a domain, while stating the meaning and properties of the domain's entities. Specifically in Life Sciences, knowledge graphs can empower hypothesis driven-experimentation [24] with knowledge extracted from the integration of various data sources or collected from community-maintained knowledge graphs (e.g., DBpedia and Wikidata) [15]. In the context of biomedicine, several knowledge graphs have been created [15, 19, 26, 37, 40, 48, 52]. They represent exemplar frameworks that put the potential of knowledge graphs into perspective by providing the knowledge required to discover novel patterns, e.g., drug-drug interactions [37, 48], cancer biomarkers [26], and cytokine levels as a biomarker [40]. Built on these results, we devise DE4LungCancer and develop a framework where stakeholders can share biomedical data sources which are integrated into the DE4LungCancer KG. Contrary to the previously mentioned approaches, DE4LungCancer relies on mapping languages, i.e., RML and FnO, to specify the process of knowledge graph creation declaratively. DE4LungCancer provides Web services to traverse the DE4LungCancer mapping rules and the unified schema, enhancing, thus, the transparency of the data integration tasks. Although the DE4LungCancer KG integrates clinical data of lung cancer patients, the DE4LungCancer framework implements the query processing methods proposed by Endris et al. [13] and ensures that data privacy regulations are respected during the execution of queries against the DE4LungCancer KG.

Quality & Ethics-Aware Data Management: To ensure data validity, and address ethical considerations and security risks of the Electronic Health Record use in such Ecosystems, the best practices have to be followed concerning data integrity, privacy and security [5]. KnowLife [14] presents an annotation-based error analysis method, in order to assess the quality of automatic construction of knowledge bases from unstructured online sources, such as the biomedical literature. Authors in [42] construct a diseases-symptoms KG from electronicistic models, and evaluate its quality by comparing to the Google health KG. A most relevant and holistic approach is presented in [20], where authors analyze the data management, legal and ethical requirements posed in big data ecosystems and illustrate a knowledge-driven architecture, in order to fulfill those. Built on these results, DE4LungCancer implements data management principles and respects the guidelines stated in [20] to provide a high-quality and ethics-aware knowledge-driven framework capable of answering clinical research questions.

7. Conclusions and Future Work

In this paper, the protagonist role of knowledge-driven data ecosystems (DEs) for enhancing transparency has been discussed. DE4LungCancer has been presented as the computational framework to address the data management, clinical, and ethical and legal requirements of the lung cancer pilot of the H2020 EU projects BigMedilytics and CLARIFY, and in the EraMed project P4-LUCAT. DE4LungCancer is a nested framework that comprises three DEs that process and analyzes the pilot datasets. DE4LungCancer offers a semantic layer composed of a unified schema, biomedical ontologies, and mapping languages; they provide the basis for a transparent data integration process into a KG. The hybrid approach that combines the pilot multi-disciplinary team with computational tools to validate integrity constraints, the unified schema, and the mapping rules have enhanced the trustability of the outcomes of the analytical services. More importantly, the documentation of the whole process backups the certification of the process by ethical committees and data protection officers. The project clinical partners can access the

DE4LungCancer services through a dashboard. the outcome of the execution of the provided services have enhanced the understanding of the conditions of the hospital services visited by lung cancer patients. Based on the observed results, clinical interventions have been devised. We plan to develop analytical methods to analyze the results of the interventions towards the improvement of the patients' quality of life.

Acknowledgements

This work has been supported by the EU H2020 funded projects BigMedilytics (GA No. 780495), the EraMed project P4-LUCAT (GA No. 53000015), and the EU H2020 RIA project CLARIFY (GA No. 875160). Furthermore, Maria-Esther Vidal is partially supported by Leibniz Association in the program "Leibniz Best Minds: Programme for Women Professors", project TrustKG-Transforming Data in Trustable Insights with grant P99/2020.

References

[1] J. Aaen, J. A. Nielsen, and A. Carugati. The dark side of data ecosystems: A longitudinal study of the damd project. *European Journal of Information Systems*, pages 1–25, 2021.

[2] M. Acosta, E. Simperl, F. Flöck, and M. Vidal. Enhancing answer completeness of SPARQL queries via crowdsourcing. *J. Web Semant.*, 45:41–62, 2017.

[3] A. R. Aronson. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*, 1:26, 2006.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of ISWC + ASWC*, pages 722–735, 2007.

[5] E. A. Balas, M. M. Vernon, F. Magrabi, L. T. Gordon, J. Sexton, et al. Big data clinical research: validity, ethics, and regulation. In *MedInfo*, pages 448–452, 2015.

[6] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[7] K. Bougiatiotis, F. Aisopos, A. Nentidis, A. Krithara, and G. Paliouras. Drug-drug interaction prediction on a biomedical literature knowledge graph. In *International Conference on Artificial Intelligence in Medicine*, pages 122–132. Springer, 2020.

[8] K. Bougiatiotis, R. Fasoulis, F. Aisopos, A. Nentidis, and G. Paliouras. Guiding graph embeddings using path-ranking methods for error detection innoisy knowledge graphs. *arXiv preprint arXiv:2002.08762*, 2020.

[9] À. Bravo Serrano, J. Piñero González, N. Queralt Rosinach, M. Rautschka, and L. I. Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics. 2015 Feb 21; 16 (1): 55*, 2015.

[10] D. S. Char, N. H. Shah, and D. Magnus. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11):981, 2018.

[11] H. Chen, S. S. Fuller, C. Friedman, and W. Hersh. Knowledge management, data mining, and text mining in medical informatics. In *Medical Informatics*, pages 3–33. Springer, 2005.

[12] A. Dimou, M. V. Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. V. de Walle. RML: A generic language for integrated RDF mappings of heterogeneous data. In *Proceedings of the Workshop on Linked Data on the Web co-located with WWW*, 2014.

[13] K. M. Endris, Z. Almhithawi, I. Lytra, M. Vidal, and S. Auer. BOUNCER: privacy-aware query processing over federations of RDF datasets. In S. Hartmann, H. Ma, A. Hameurlain, G. Pernul, and R. R. Wagner, editors, *Database and Expert Systems Applications - 29th International Conference, DEXA 2018, Regensburg, Germany, September 3-6, 2018, Proceedings, Part I*, volume 11029 of *Lecture Notes in Computer Science*, pages 69–84. Springer, 2018.

[14] P. Ernst, A. Siu, and G. Weikum. Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC bioinformatics*, 16(1):1–13, 2015.

[15] A. W. et al. Wikidata as a knowledge graph for the life sciences. *Science Forum eLife 2020;9:e52614*, 2020.

[16] M. P. et all. Neoadjuvant chemotherapy and nivolumab in resectable non-small-cell lung cancer (nadim): an open-label, multicentre, single-arm, phase 2 trial. *The Lancet Oncology*, 2020.

[17] Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, 2018.

[18] M. Figuera, P. D. Rohde, and M. Vidal. Trav-shacl: Efficiently validating networks of SHACL constraints. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3337–3348, 2021.

[19] M. Färber and D. Lamprecht. The data set knowledge graph: Creating a linked open data source for data sets. *Quantitative Science Studies*, 2(4):1324–1355, 2021.

[20] S. Geisler, M. Vidal, C. Cappiello, B. F. Lóscio, A. Gal, M. Jarke, M. Lenzerini, P. Missier, B. Otto, E. Paja, B. Pernici, and J. Rehof. Knowledge-driven data ecosystems toward data transparency. *ACM J. Data Inf. Qual.*, 14(1):3:1–3:12, 2022.

[21] P. Groth and M. Dumontier. Introduction - FAIR data, systems and analysis. *Data Sci.*, 3(1):1–2, 2020.

[22] C. Gutiérrez and J. F. Sequeda. Knowledge graphs. *Commun. ACM*, 64(3):96–104, 2021.

[23] L. Halilaj, N. Petersen, I. Grangel-González, C. Lange, S. Auer, G. Coskun, and S. Lohmann. Vocol: An integrated environment to support version-controlled vocabulary development. In *Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings*, pages 303–319, 2016.

[24] T. Hulsen, S. S. Jamuar, A. R. Moody, J. H. Karnes, O. Varga, S. Hedensted, R. Spreafico, D. A. Hafler, and E. F. McKinney. From big data to precision medicine. *Frontiers in Medicine*, 6, 2019.

[25] E. Iglesias, S. Jozashoori, D. Chaves-Fraga, D. Collarana, and M.-E. Vidal. Sdm-rdfizer: An rml interpreter for the efficient creation of rdf knowledge graphs. In *ACM International Conference on Information & Knowledge Management*, 2020.

[26] A. Jha, Y. Khan, M. Mehdi, M. R. Karim, Q. Mehmood, A. Zappa, D. Rebholz-Schuhmann, and R. Sahay. Towards precision medicine: discovering novel gynecological cancer biomarkers and pathways using linked data. *J. Biomed. Semant.*, 8(1):40:1–40:16, 2017.

[27] S. Jozashoori, D. Chaves-Fraga, E. Iglesias, M. Vidal, and Ó. Corcho. Funmap: Efficient execution of functional mappings for knowledge graph creation. In *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference*, 2020.

[28] S. Jozashoori, A. Sakor, E. Iglesias, and M. Vidal. Eablock: A declarative entity alignment block for knowledge graph creation pipelines. In *The ACM Symposium on Applied Computing, SAC*, 2022.

[29] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, 2012.

[30] H. Knublauch and D. Kontokostas. Shapes constraint language (shacl). W3C Recommendation, 2017.

[31] N. Lao and W. W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81(1):53–67, 2010.

[32] C. H. Lee and H.-J. Yoon. Medical big data: promise and challenges. *Kidney research and clinical practice*, 36(1):3, 2017.

[33] S. Marjanovic, I. Ghiga, M. Yang, and A. Knack. Understanding value in health data ecosystems: A review of current evidence and ways forward. *Rand health quarterly*, 7(2), 2018.

[34] A. Melo and H. Paulheim. Detection of relation assertion errors in knowledge graphs. In *Proceedings of the Knowledge Capture Conference*, pages 1–8, 2017.

[35] G. A. Mihaila, L. Raschid, and M. Vidal. Using quality of data metadata for source selection and ranking. In *Proceedings of the Third International Workshop on the Web and Databases, WebDB 2000, Adam's Mark Hotel, Dallas, Texas, USA, May 18-19, 2000, in conjunction with ACM PODS/SIGMOD 2000. Informal proceedings*, pages 93–98, 2000.

[36] B. D. Mittelstadt and L. Floridi. The ethics of big data: current and foreseeable issues in biomedical contexts. *The ethics of biomedical big data*, pages 445–480, 2016.

[37] D. N.Nicholson and C. S. Greene. Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, 18:1414–1428, 2020.

[38] M. I. S. Oliveira, G. d. F. B. Lima, and B. F. Lóscio. Investigations into data ecosystems: a systematic mapping study. *Knowledge and Information Systems*, 61(2):589–630, 2019.

[39] E. Prud'hommeaux and A. Seaborne. Sparql query language for rdf. W3C Recommendation, 2008.

[40] N. Queralt-Rosinach, R. Kaliyaperumal, and e. a. C.H. Bernabe. Applying the fair principles to data in a hospital: challenges and opportunities in a pandemic. *J Biomedical Semantics*, 13(12), 2022.

[41] T. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, 2003.

[42] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag. Learning a health knowledge graph from electronic medical records. *Scientific reports*, 7(1):1–11, 2017.

[43] E. Ruckhaus, M. Vidal, S. Castillo, O. Burguillos, and O. Baldizan. Analyzing linked data quality with liquate. In *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, pages 488–493, 2014.

[44] A. Sakor, I. O. Mulang, K. Singh, S. Shekarpour, M. Vidal, J. Lehmann, and S. Auer. Old is gold: Linguistic driven approach for entity and relation linking of short text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2336–2346, 2019.

[45] A. Sakor, K. Singh, A. Patel, and M. Vidal. Falcon 2.0: An entity and relation linking tool over wikidata. In *The 29th ACM International Conference on Information and Knowledge Management- CIKM*, 2020.

[46] M. Scurti, E. M. Ruiz, M. Vidal, M. Torrente, D. Vogiatzis, G. Paliouras, M. Provencio, and A. R. González. A data-driven approach for analyzing healthcare services extracted from clinical records. In *33rd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2020*, 2020.

[47] E. Vayena, A. Blasimme, and I. G. Cohen. Machine learning in medicine: addressing ethical challenges. *PLoS medicine*, 15(11):e1002689, 2018.

[48] M. Vidal, K. M. Endris, S. Jazashoori, A. Sakor, and A. Rivas. Transforming heterogeneous data into knowledge for personalized treatments - A use case. *Datenbank-Spektrum*, 19(2):95–106, 2019.

[49] D. Vrandecic and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

[50] M. Wiesenauer, C. Johner, and R. Röhrig. Secondary use of clinical data in healthcare providers–an overview on research, regulatory and ethical requirements. *Quality of Life through Quality of Information*, pages 614–618, 2012.

[51] R. Wood and G. Taylor-Stokes. Cost burden associated with advanced non-small cell lung cancer in europe and influence of disease stage. *BMC cancer*, 19(1), 2019.

[52] J. Yuan, Z. Jin, H. Guo, H. Jin, X. Zhang, T. H. Smith, and J. Luo. Constructing biomedical domain-specific knowledge graph with minimum supervision. *Knowl. Inf. Syst.*, 62(1):317–336, 2020.

[53] S. Zheng, J. Rao, Y. Song, J. Zhang, X. Xiao, E. F. Fang, Y. Yang, and Z. Niu. Pharmkg: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in Bioinformatics*, 22(4):bbaa344, 2021.